6

Perceptive characteristics of speech of untrained and trained subjects, and influences of gender and age

Arend M. Sulter and Herman F. Peters*

Department of Otorhinolaryngology, University Hospital Groningen, Groningen, and * Department of Otorhinolaryngology, University Hospital Nijmegen, Nijmegen, the Netherlands

Submitted

INTRODUCTION

This study is the result of an extensive research project, originated to supply information on normal voice production and to create a frame of reference for qualifying vocal performance. Apart from a large group of subjects without vocal complaints or vocal pathology, another group of subjects having received vocal training and regularly exploiting their vocal abilities, was investigated by different means, to give direction to what should be regarded as good vocal performance in the continuum from poor to excellent. Related studies have already established normative data regarding phonetograms (1), laryngostroboscopy (2), and glottal closure (3). In this article speech characteristics of subjects are evaluated.

The complex acoustic event representing speech results from a basic voice source signal and the modulation of this signal by the articulatory organs. One important method to qualitatively analyze speech is perceptual evaluation (4). Although listeners vary in their qualitative description of speech sounds (5-7), a structured approach using a multidimensional scaling instrument provides reliable perceptual judgments on running speech with constantly emerging factorial dimensions (8,9).

Previous studies resulting from our research project concentrated on

characteristics of the voice source. Differences between men and women were observed in melodic and intensity ranges (1), in both laryngeal appearance and glottal functioning (2,3), as well as in vocal fold physiology (4). Less appealing differences were noted between trained and untrained subjects; however, results of phonetography indicated that trained subjects might benefit from a superior control over the voice source (1).

Speech has been the subject of many investigations, concentrating on different aspects, amongst others perception of speech. However, limited information is available on speech characteristics, as determined with a standardized scaling instrument in large groups of subjects. Using a standardized scaling instrument has several advantages. It employs the tedious work of determining scales that can be utilized in a rating experiment. An instrument with selected scales forces judges to express their opinion in a standardized way, thus excluding incompatible variety of expressions. Using a specific scaling instrument also offers the possibility of comparing results between investigations. One of the few studies using a standardized scaling instrument on evaluating laryngeal speech of large groups of men and women was conducted by Tielen (9). Only a few distinct differences were determined between men and women. Compared to men, women were evaluated to speak with higher pitch and more melodious. The specific choice of subjects, in relation with their profession, might have caused the resemblance in speech characteristics between men and women. Other cohorts of men and women might have revealed a different picture, more close to what is expected from literature. The scaling instrument employed by Tielen (9) was modified to concentrate on differences in sociocultural aspects of speech. Scales more specifically reflecting physiology of phonation had a less prominent place, although this information could be used to relate perceptual aspects of speech with a physiologic basis to both quality of speech and physiologic measurements.

In previous studies trained subjects showed a possible superior control over the voice source (1,11). Perceptual evaluation of speech of trained subjects might demonstrate effects of this superior control.

With the former considerations, the present study investigates aspects of speech and related vocal and articulatory processes, to answer the following questions:

1. Can speech of a large number of voice healthy subjects reliably be evaluated with the perceptual scaling instrument of Fagel et al. (8)?

2. Does factor analysis of scale scores result in a practicable solution, fit for a further evaluation of perceptive data?

3. Do groups differ in scale scores; and if there are differences, how can they be specified according to the grouping variables gender and vocal

training?

4. Are differences between groups also reflected in differences in factor scores and what do they represent in perceptual dimensions?

5. Which scales do best differentiate between men and women, and between subjects with and without vocal training?

METHODS

Nieboer, De Graaf & Schutte (12) evaluated speech of alaryngeal speakers in a similar way we intended to do for laryngeal speech in the present study, also using --a modified version of-- the scaling instrument proposed by Fagel et al. (8). Therefore the procedural approach of Nieboer et al. (12) was used as a guideline.

Subjects

Speech samples were provided by a total of 224 Dutch untrained and trained subjects of both genders, categorized accordingly into 4 groups. The untrained subjects were recruited from groups of students and volunteers without vocal complaints or history of vocal pathology. The group consisted of 92 females and 47 males. The mean age for the female subjects in this subgroup was 20.3 years, ranging from 17 to 44 (median 19 years; standard deviation [SD] 7.37), while the mean age for the male subjects was 25.0 years, ranging from 17 to 35 (median 25 years; SD 4.68 years). Eighteen of the female, and 16 of the male subjects were smokers.

42 female and 43 male amateur singers with a minimum of two years of vocal training served as another group. The vocal training could either consist of singing in a choir that organized rehearsals with a minimum frequency of once a week, or receiving individual singing lessons with a similar minimum frequency. All choirs had a professional conductor and used auditions to admit new members. Although a minimum of 2 years of organized singing was used as a selection criterion to be included in the trained group, about 60% of the trained subjects had a considerably longer history of singing in a choir (> 5 years). The mean age of the female trained group was 35.1 years, ranging from 18 to 59 (median 34 years; SD 11.86 years), and the mean age of the male subjects was 47.5 years, ranging from 21 to 75 (median 49 years; SD 18.52 years). Five of the female, and 11 of the male trained subjects were smokers. Because all participants in this study volunteered, we refrained from matching according to age.

Scales

Speech characteristics were analyzed by way of 14 bipolar semantic scales

with seven points (13), ranging from -3 to 3. The scales were taken from Fagel et al. (8), who carefully developed a scaling instrument by reducing numerous adjectives to a practicable number of so-called Alpha scales, making possible a global perceptual description of a speaker in a multidimensional perceptual space. A few changes were implemented. The scale "creaky-not creaky" was introduced to focus on a voice characteristic not incorporated in the original instrument, and the scales "dragging-brisk" and "slow-quick" were replaced by the alternative "slurring-sprightly". Instead of "husky-not husky" the scale ends "breathy-not breathy" will be used, because they reflect more closely the original Dutch terms.

The scales were all given the same direction or polarization: the more "negative" or "unfavourable" pole was placed on the left-hand side, the more "positive" or "favourable" pole on the right-hand side.

Adjectives of the scales were translated from Dutch. Minor shifts in meaning may therefore have occurred. The Dutch and English terms are listed in Table 1. Throughout this paper, English terms will be used. While referring to the results of this study, the reader is advised to give the original Dutch term with the English translation.

Listeners were also asked to estimate the age of the speaker. This information might be used to study correlations with other scales.

Scale	factor 1		factor 2		factor 3	
	eigen-value	% var	eigen-value	% var	eigen-value	% var
1 expressionless-expressive (expressieloos-expressief)	12.3	45.7	1.5	5.5	1.3	5.0
2 monotonous-melodious (monotoon-melodieus)	12.0	44.6	1.5	5.7	1.4	5.0
3 slurring-sprightly (slepend-levendig)	11.2	41.5	1.5	5.7	1.3	4.9
4 shrill-warm (schel-warm)	12.5	46.2	1.4	5.2	0.9	3.5
5 high-low for a (wo)man (hoog-laag voor een man/vrouw)	12.3	45.6	1.2	4.5	1.1	4.2
6 ugly-beautiful (lelijk-mooi)	12.1	44.7	1.7	6.1	1.1	4.0
7 unpleasant-pleasant (onplezierig-plezierig)	11.9	43.9	1.7	6.4	1.3	4.6
8 breathy-not breathy (hees-niet hees)	11.5	42.6	1.4	5.3	1.2	4.3
9 creaky-not creaky (krakerig-niet krakerig)	8.1	30.1	1.7	6.2	1.4	5.3
10 dull-clear (dof-helder)	8.5	31.4	1.5	5.7	1.5	5.5
11 soft-loud (zacht-luid)	10.5	38.8	1.4	5.0	1.3	4.8
12 weak-powerful (zwak-krachtig)	10.2	37.9	1.3	4.9	1.3	4.6
13 broad-cultured (plat-beschaafd)	10.8	39.9	1.4	5.1	1.3	4.8
14 slovenly-polished (slordig-netjes)	9.3	34.5	1.6	6.0	1.4	5.3
15 estimated age (geschatte leeftijd)	22.1	81.9	0.5	1.9	0.5	1.7

Table 1. Factor analysis of the scores on 15 scales given by 27 listener judges. % var = percentage of variation. Scales are given in the original Dutch version (italics) and an English translation.

Speech samples

Speech stimuli were obtained by making high quality recordings of subjects reading a text of neutral content (De noorderwind en de zon). Equipment

consisted of a B&K 4003 microphone, a B&K amplifier (type 2812) and a Sony PCM SL-F1E recorder. Recording level was adjusted for each subject to optimize signal to noise ratio. A text was preferred above spontaneous speech to control for individually differing lexicon and syntax. From the recordings three stimulus tapes were made, containing the text of about 50 seconds of speech, hereafter referred to as a sample, of each subject. Samples were randomly copied to these three tapes. The first tape started with 5 samples to have material for the judges to practise rating the scales. These scores were not used for further evaluation. The next 10 samples were also used as the last samples on the third tape in order to provide information on intrajudge reliability.

Listeners

Twenty-seven female students of the Academy for Logopedics in Nijmegen (mean age 25.0 years, standard deviation 6.07 years) in the third year of the training course rated all samples with the scaling instrument discussed. The students were regarded as naive judges because of both the fact that during their training little time was spend on perceptual evaluation of voices, as well as the limited time available to score each scale (about three seconds). Naive judges were favoured, as they are known to give adequate judgments (12) and more uniform ratings than expert raters (5), as well as time problems with expert raters, given the large numbers of samples to be judged. The presence of only female judges should not result in biased ratings (8,9). Each of the judges was given an honorarium of hfl. 40,-.

Rating procedure

The judges were instructed to rate according their first impression of the speech. To make them familiar with using the scaling instrument they were presented a randomly chosen number of five speakers.

All samples were rated in three afternoon sessions of 2 hours, each on the same day of consecutive weeks. Each session consisted of three blocks of 30 minutes rating time with in-between breaks of 10 minutes. None of the judges reported loss of concentration during the experiment.

Filtering and amplification of the sound were adjusted in such a way as to resemble as much as possible the unfiltered sound as it could be heard using headphones. The sound quality in different places in the room was checked both before and during the experiment. Care was taken to give all speech samples approximately the same loudness level when they were played to the listeners.

Data reduction

Descriptive and inferential statistical evaluation was performed with the SPSS software package (SPSS Inc.). Means and standard deviations per speaker per scale, as well as bar charts of the scores were computed. Two-way Analysis of (Co)Variance (ANCOVA) was used to determine significant effects of the grouping variables gender and vocal training, as well as the covariant age on scales and factors. If a significant interaction was present between grouping variables, separate oneway ANOVAs were performed for each grouping variable. The significance level was set at $p=0.05$.

Factor analysis was performed on the scores in order to gain insight into the dimensionality of the perceptual space used for judging the speakers. The factor analysis performed was a principal-component analysis with iteration. A factor solution with 6 components was pursued, five of the components being reserved for the solution found by Fagel et al. (8), and one for the estimated age. Factors with an eigenvalue < 1.0 were therefore accepted. The factors produced in this way were rotated according to the varimax procedure in order to get a clearer factor configuration (14,15). The varimax rotation does not affect the orthogonal factor structure. This means that factors are independent of each other.

For each speaker, factor scores on each factor were computed by multiplying the speaker's standardized mean score on those scales loading highest on that factor, with the corresponding factor-score coefficient. The sum of these products is the factor score. As six main factors were selected, each speaker could be characterized with the six figures representing his/her factor scores. Factors scores are calculated on the basis of the standardized scores; therefore the factor scores can be either positive or negative.

A discriminant analysis was performed on the scale means of the speakers, in order to determine which set of scales could best discriminate between men and women, and, separately for each gender, between untrained and trained subjects. The analysis was carried out according to the Rao's V method, which, in the creation of a discriminant function, selects or deletes variables on the basis of their contribution to the increase in Rao's V. Rao's V is a generalized measure of the distance between the groups along the one possible dimension.

RESULTS

Scales properties

The uniformity in using a same definition of scale ends by the listener group was checked by performing a factor analysis on all scores given by the listeners, separately for each scale. Table 1 gives the result of the factor analysis. The first three factors are given to show the structure of the factor solution. All scales have by far the highest loading on the first factor. Apart

from estimated age, eigenvalues range from 12.5 to 8.1 for scales "shrill-warm" and "creaky-not creaky", respectively. The second factor shows much smaller eigenvalues ranging from 1.7 to 1.2. Therefore ratings given on scales can be regarded as given by one group with a same representation of scale ends and thus of scale use.

Table 2 shows mean correlation between raters, effective reliabilities, and minimum number of raters needed to obtain an effective reliability of 0.90 and 0.95, for the 15 scales.

$$R_e = \frac{nr_m}{(1+(n-1)r_m)} \qquad (1)$$

The "effective reliability" or "standardized item alpha" (where "item" is to be read in our case as "rater", SPSS Inc., (16)) of the 15 scales used in the rating experiment was computed according to the formula (1) where R_e is the effective reliability, n is the number of raters, and r is the mean correlation between raters. In general, high effective reliability figures were found. Values are above 0.9, with the exception of the scales "creaky-not creaky" and "dull-clear", which show reliability values of 0.884 and 0.895, respectively.

In general, ratings on a scale with an effective reliability of > 0.90 are

	Scale	Mean correlation between raters	Effective reliability	Minimum number of raters for effective reliability of 0.90	Minimum number of raters for effective reliability of 0.95
1	expressionless-expressive	0.40	0.947	14	29
2	monotonous-melodious	0.39	0.946	15	30
3	slurring-sprightly	0.35	0.935	17	36
4	shrill-warm	0.39	0.944	15	30
5	high-low for a (wo)man	0.35	0.936	17	36
6	ugly-beautiful	0.38	0.943	15	31
7	unpleasant-pleasant	0.36	0.939	16	34
8	breathy-not breathy	0.35	0.937	17	36
9	creaky-not creaky	0.22	0.884	32	68
10	dull-clear	0.24	0.895	29	61
11	soft-loud	0.33	0.931	19	39
12	weak-powerful	0.32	0.928	20	41
13	broad-cultured	0.33	0.930	19	39
14	slovenly-polished	0.26	0.907	26	55
15	estimated age	0.78	0.990	3	6

Table 2. Mean correlation between raters, effective reliability, and minimum number of raters needed to obtain an effective reliability of 0.90 and 0.95, for the 15 scales used in the rating experiment

considered to give reliable information (17). The minimum number of raters needed to obtain a pre-defined effective reliability is calculated according to the formula

$$n_{min} = \frac{1 - r_m \quad R_e}{r_m \quad 1 - R_e} \qquad (2)$$

where n_{min} is the minimum number of judges, R_e is the effective reliability to be obtained, and r_m is the mean correlation between raters. For the bipolar semantic scales the minimum number of raters needed to obtain an effective reliability of 0.90 ranged from 14 ("expressionless-expressive") to 32 ("creaky-

	Scale	c.c.	p-level
1	expressionless-expressive	0.91	$p < 0.001$
2	monotonous-melodious	0.89	$p < 0.001$
3	slurring-sprightly	0.93	$p < 0.001$
4	shrill-warm	0.89	$p = 0.001$
5	high-low for a (wo)man	0.91	$p < 0.001$
6	ugly-beautiful	0.92	$p < 0.001$
7	unpleasant-pleasant	0.92	$p < 0.001$
8	breathy-not breathy	0.91	$p < 0.001$
9	creaky-not creaky	0.95	$p < 0.001$
10	dull-clear	0.87	$p = 0.001$
11	soft-loud	0.91	$p < 0.001$
12	weak-powerful	0.92	$p < 0.001$
13	broad-cultured	0.89	$p = 0.001$
14	slovenly-polished	0.69	$p = 0.028$
15	estimated age	0.99	$p < 0.001$

Table 3. Averaged intrajudge correlation coefficients (c.c.) and corresponding probability level, based on the two ratings (test-retest) the 27 listener judges gave on 10 speech samples.

not creaky"). The small number of three raters is already enough to present a reliable --not necessarily valid-- estimation of age.

Table 3 gives information on intrajudge reliability for the scales. Values are based on calculating a correlation coefficient between scores on scales at the beginning and at the end of the rating experiment. Only mean values are given, as the raters could be regarded as one group. Apart from the value for the scale "slovenly-polished" (0.69), which comes as an extreme compared to other figures, correlation coefficients range from 0.87 ("dull-clear") to 0.95 ("creaky-not creaky") for the bipolar semantic scales. Estimated age shows a high value of 0.99. Generally, it can be concluded that no change in using the scaling instrument occurred over the experiment as a whole.

Mean scores

Table 4 gives calculated mean values and standard deviations for the 92 untrained and 42 trained female, and 47 untrained and 43 trained male subjects on each scale. With a few exceptions, mean values range between scale values -1 and 1 with standard deviations ranging from 0.6 to 1.0. Compared to untrained females, trained ones show higher averaged values on all scales, that is, their speech is rated more positively on an averaged base. Compared to untrained males, trained males have both higher and lower averaged values. Speech of trained males is especially rated more breathy, creaky, dull and broad on an averaged base.

No debate about gender will be held here, because of the inconsistent

differences in mean scale values, as well as lacking information on statistical significance of these differences (see next section for significance of differences).

Except for untrained men, mean estimated ages are within 4 years of the correct mean age, which shows, on an averaged base, the good impression we get of age by listening to speech of a person.

Group differences in mean scores

To analyze significant differences between men and women, and untrained and trained subjects ANCOVA was performed. Table 5 gives the results of these analyses. With the information on mean scale values from table 4 the following observations can be made: Regarding gender, speech of women was rated more expressive ($p<0.001$) and melodious ($p<0.001$), higher ($p<0.001$), and clearer ($p=0.032$) on the one hand, and more unpleasant ($p=0.005$), softer ($p=0.037$), and weaker ($p<0.001$) on the other.

	Scale	Women		Men	
		Untrai-ned	Trained	Untrai-ned	Trained
1	expressionless-expressi-ve	0.68 (0.81)	1.13 (0.70)	0.30 (0.86)	0.58 (0.98)
2	monotonous-melodious	0.81 (0.77)	1.24 (0.67)	0.38 (0.84)	0.59 (0.93)
3	slurring-sprightly	0.41 (0.72)	0.58 (0.78)	0.24 (0.83)	0.27 (0.92)
4	shrill-warm	-0.22 (0.68)	0.36 (0.78)	1.00 (0.65)	1.03 (0.63)
5	high-low for a (wo)man	-0.57 (0.59)	-0.28 (0.59)	0.44 (0.49)	0.52 (0.60)
6	ugly-beautiful	-0.23 (0.74)	0.35 (0.81)	0.42 (0.87)	0.32 (0.93)
7	unpleasant-pleasant	-0.20 (0.77)	0.41 (0.83)	0.34 (1.00)	0.42 (1.03)
8	breathy-not breathy	0.38 (0.90)	0.87 (0.74)	0.87 (0.65)	0.69 (1.01)
9	creaky-not creaky	0.59 (0.65)	0.65 (0.77)	0.48 (0.72)	0.26 (0.80)
10	dull-clear	0.57 (0.63)	0.80 (0.64)	0.42 (0.64)	0.33 (0.80)
11	soft-loud	0.36 (0.67)	0.36 (0.64)	0.51 (0.64)	0.67 (0.58)
12	weak-powerful	0.10 (0.67)	0.31 (0.63)	0.55 (0.68)	0.64 (0.67)
13	broad-cultured	0.25 (0.87)	0.86 (0.80)	1.03 (0.60)	0.32 (0.91)
14	slovenly-polished	0.43 (0.78)	1.22 (0.69)	0.58 (0.82)	0.77 (0.70)
15	estimated age	23.72 (4.39)	32.98 (8.39)	31.37 (3.16)	44.56 (8.96)

Table 4. Mean score values and standard deviations (between brackets) of groups on the 15 scales.

	Scale	Gender		Vocal training		Gender x Training		Age		
		F	p	F	p	F	p	F	p	rrc
1	expressionless-expressive	13.79	<0.001*	5.80	0.017*	0.57	0.449	0.53	0.470	0.003
2	monotonous-melodious	18.86	<0.001*	5.64	0.018*	0.87	0.352	0.04	0.836	0.001
3	slurring-sprightly	2.68	0.103	1.79	0.183	0.23	0.630	0.60	0.438	-0.003
4	shrill-warm	101.73	<0.001*	10.30	0.002*	7.56	0.006*	30.64	<0.001*	0.018
	untrained/women resp.	100.59	<0.0001*	19.00	<0.0001*					
	trained/men resp.	18.81	<0.0001*	0.04	0.843					
5	high-low for a wo(man)	115.11	<0.001*	2.30	0.131	1.94	0.165	37.51	<0.001*	0.016
6	ugly-beautiful	13.47	<0.001*	8.58	0.004*	7.33	0.007*	2.87	0.092	0.006
	untrained/women resp.	21.65	<0.0001*	16.90	<0.001*					
	trained/men resp.	0.02	0.891	0.29	0.593					
7	unpleasant-pleasant	8.00	0.005*	8.87	0.003*	3.78	0.053	5.22	0.023*	0.009
8	breathy-not breathy	4.41	0.037*	3.60	0.059	7.17	0.008*	1.15	0.284	0.004
	untrained/women resp.	11.15	0.001*	9.51	0.003*					
	trained/men resp.	0.85	0.358	1.07	0.303					
9	creaky-not creaky	2.51	0.115	0.33	0.566	1.15	0.285	6.03	0.015*	-0.008
10	dull-clear	4.64	0.032*	4.72	0.031*	1.67	0.198	4.53	0.034*	-0.007
11	soft-loud	4.39	0.037*	0.02	0.883	0.57	0.450	3.60	0.059	0.006
12	weak-powerful	16.65	<0.001*	1.63	0.203	0.43	0.511	7.89	0.005*	0.009

Table 5. Analysis of (co)variance summary table with effects of grouping variables and covariant on scales.
df = 1 , 222 for gender, vocal training, and gender x training. rrc, raw regression coefficient. * $p < 0.05$.

Regarding vocal training, speech of trained subjects was rated more expressive ($p=0.017$), melodious ($p=0.018$) and pleasant ($p=0.003$), and also clearer ($p=0.031$).

A significant interaction between grouping variables gender and vocal training was found for the scales "shrill-warm" ($p=0.006$), "ugly-beautiful" ($p=0.007$), "breathy-not breathy" ($p=0.008$), "broad-cultured" ($p<0.001$), and "slovenly-polished" ($p=0.004$). Therefore, separate oneway ANOVAs were performed at each grouping variable level. Speech of untrained women was rated more shrill ($p<0.0001$), ugly ($p<0.0001$), breathy ($p=0.001$) and broad ($p<0.0001$), as compared to speech of untrained men. Compared to trained men, speech of trained women was also rated more shrill ($p<0.0001$), however, it was rated less broad ($p=0.005$) and slovenly ($p=0.003$). Speech of untrained women was rated more shrill ($p<0.0001$), ugly ($p<0.0001$), breathy ($p=0.003$), broad ($p<0.001$) and slovenly ($p<0.0001$), as compared to speech of trained women. Finally, speech of trained men was rated more broad ($p<0.0001$), as compared to speech of untrained men.

Age of the speaker had on the one hand a significantly positive influence on the scales "shrill-warm" ($p<0.001$), "high-low" ($p<0.001$), "unpleasant-pleasant" ($p=0.023$), "weak-powerful" ($p=0.005$) and "slovenly-polished" ($p=0.003$), and a significantly negative influence on the scales "creaky-not creaky" ($p=0.015$) and "dull-clear" ($p=0.034$) on the other.

Estimated age was rated significantly different for the grouping variables gender ($p<0.001$) and vocal training ($p=0.001$), men being older than women, and trained subjects being older than untrained ones. Age had a highly positive significant influence on estimated age ($F(1,222)=1227.06$; $p<0.001$).

Factor loadings

Table 6 shows the loadings, after varimax rotation, of each of the 15 scales on the six main factors extracted by means of factor analysis.

A clear factor structure emerges from the analysis. The percentage of the variance accounted for by six factors was 93.0. The eigenvalues of the six factors (i.e. the summation of the squared loadings on the 15 scales on each factor, expressing the amount of variation in the scales explained by that factor) after varimax rotation were: factor 1, 6.11; factor 2, 2.90; factor 3, 1.99; factor 4, 1.60; factor 5, 0.81; factor 6, 0.54. The communality h^2 (i.e. the summation of the squared loadings of one scale on each of the factors) ranged from 0.83 ("creaky-not creaky") to 0.97 ("expressionless-expressive" and "weak-powerful").

A first factor emerged with high loadings on the scales "expressionless-expressive", "monotonous-melodious" and "slurring-sprightly". Because all three scales are related to prosodic features of speech, this factor was given the

term Intonation.

The second factor was a combined evaluative factor, with a strong component representing pitch level by the scales "shrill-warm" and "high-low for a wo(man)" on the one hand, and a qualitative component represented by the scales "ugly-beautiful" and "unpleasant-pleasant" on the other. As the scale "shrill-warm" also represents an emotional impression of speech, this factor was given the term Quality.

In the third factor high loadings were found on the scales "breathy-not breathy", "creaky-not creaky" and "dull-clear", all representing characteristics of the voice source. Therefore this factor was labelled Physiology.

A fourth factor was labelled Dynamics, because of the high loadings on the scales "soft-loud" and "weak-powerful".

	Scale	Loadings of the scale on the factor						h
		1:Intonati-on	2:Quality	3:Physiology	4:Dynamics	5:Articulation	6:Estimated age	
1	expressionless-expressi-ve	0.93	-0.05	0.10	0.19	0.22	0.03	0.97
2	monotonous-melodious	0.92	-0.10	0.13	0.18	0.23	0.00	0.96
3	slurring-sprightly	0.91	0.04	0.16	0.23	0.09	-0.11	0.93
4	shrill-warm	-0.05	0.92	0.12	0.07	0.25	0.14	0.95
5	high-low for a wo(man)	-0.33	0.86	-0.12	0.18	0.04	0.21	0.94
6	ugly-beautiful	0.39	0.64	0.46	0.12	0.40	0.01	0.95
7	unpleasant-pleasant	0.45	0.62	0.43	0.13	0.39	0.05	0.93
8	breathy-not breathy	-0.02	0.08	0.92	0.24	0.10	0.07	0.92
9	creaky-not creaky	0.17	0.19	0.79	-0.24	0.02	-0.29	0.83
10	dull-clear	0.39	-0.10	0.79	0.29	0.25	-0.08	0.94
11	soft-loud	0.25	0.06	0.07	0.93	-0.09	0.03	0.95
12	weak-powerful	0.32	0.29	0.20	0.85	-0.01	0.06	0.97
13	broad-cultured	0.21	0.28	0.03	-0.03	0.82	-0.26	0.87
14	slovenly-polished	0.27	0.19	0.24	-0.11	0.81	0.24	0.88
15	estimated age	-0.03	0.26	-0.13	0.05	-0.01	0.93	0.94
Eigenvalue		6.11	2.90	1.99	1.60	0.81	0.54	

Table 6. Factor analysis of the 14 bipolar semantic scales and estimated age scale used in the rating experiment. The rows show the scales' varimax rotated factor loadings on the six main factors, labelled Intonation, Quality, Physiology, Dynamics, Articulation and Estimated age and their communalities (h^2). The bottom rows show eigenvalues, percentage of variance accounted for, and cumulative percentage of variance accounted for by the factors. Factor loadings of 0.45 and higher are in italic.

A fifth factor was labelled Articulation, because of the high loadings on the scales "broad-cultured" and "slovenly-polished".

The scale "estimated age" was uniquely represented with a high loading on a separate factor, which therefore also was given the term Estimated age.

Group differences in factor scores

Factor scores were calculated for each of the 224 subjects. Because of the large number of subjects, no overview of individual factor scores will be given. Instead, table 7 summarizes mean values and standard deviations of the factor scores for each group. Except for factor 4, Dynamics, trained women have more positive mean values than the untrained ones, expressing the higher appreciation of their speech by the listeners. Trained men have a more positive mean value on factor 1, Intonation, while, compared to untrained men, their

	Women		Men	
Scale	Untrai-ned	Trained	Untrai-ned	Trained
Factor 1; Intonation	0.13 (0.89)	0.41 (0.86)	-0.54 (0.97)	-0.11 (1.14)
Factor 2; Quality	-0.49 (0.84)	-0.23 (1.04)	0.64 (0.84)	0.59 (0.76)
Factor 3; Physiology	-0.03 (1.01)	0.26 (0.91)	-0.04 (0.84)	-0.14 (1.19)
Factor 4; Dynamics	-0.20 (1.06)	-0.23 (0.92)	0.37 (0.93)	0.26 (0.85)
Factor 5; Articulation	-0.25 (1.00)	0.56 (0.96)	0.22 (0.82)	-0.24 (0.96)
Factor 6; Estimated age	-0.58 (0.56)	0.40 (1.03)	-0.32 (0.44)	1.19 (0.96)

Table 7. Mean factors scores and (between brackets) standard deviations.

articulation is less appreciated, regarding the more negative mean value on factor 5.

ANCOVAs were performed to analyze significant influences of grouping variables and covariant age on factor scores. Table 8 gives the results. Gender had a significant effect on factor 1, Intonation ($p<0.001$); factor 2, Quality ($p<0.001$); and factor 4, Dynamics ($p<0.001$), women having a more positively

Scale	Gender		Vocal training		Gender x Training		Age		
	F	p	F	p	F	p	F	p	rrc
Factor 1 sum score; Intonation	21.24	<0.001*	2.15	0.144	0.19	0.665	0.37	0.544	0.003
Factor 2 sum score; Quality	66.39	<0.001*	1.92	0.167	1.23	0.269	9.21	0.003*	0.012
Factor 3 sum score; Physiology	0.85	0.359	1.18	0.278	1.62	0.205	0.14	0.711	-0.002
Factor 4 sum score; Dynamics	16.03	<0.001*	0.02	0.889	0.31	0.861	0.32	0.575	0.003
Factor 5 sum score; Articulation	0.29	0.588	9.91	0.002*	18.9	<0.001*	0.47	0.492	-0.003
untrained/women resp.	7.84	0.006*	19.4	<0.0001	1				
trained/men resp.	14.58	<0.001*	0	*					
			5.98	0.017*					
Factor 6 sum score; Estimated age	1.59	0.209	9.12	0.003*	0.48	0.489	596.82	<0.001*	0.059

Table 8. Analysis of (co)variance summary table with effects of grouping variables and covariant on factor sum scores.
df = 1, 222 for gender, vocal training, and gender x training. rrc, raw regression coefficient. * p < 0.05

rated Intonation characteristic on the one hand, and more negatively rated Quality and Dynamics characteristics on the other.

A significant interaction between grouping variables was found for factor 5, Articulation ($p<0.001$). Therefore, separate oneway ANOVAs were performed at each grouping variable level. Untrained women have a significantly more negatively rated Articulation ($p=0.006$) compared with untrained men, while the opposite was observed in trained subjects ($p<0.001$). Untrained women have a significantly more negatively rated Articulation ($p<0.0001$) compared with trained women, while Articulation characteristic of speech of untrained men is rated significantly more positive ($p=0.017$), as compared to trained men.

Although significant influences of vocal training were found on scales, no specific influences were observed on calculated factor scores. Estimated age being the only scale present in factor 6, again showed to be significantly influenced by vocal training ($p=0.003$), untrained subjects having a younger rated age.

Age as a covariant had a significantly positive influence on factor 2, Quality, implying that speech of older subjects is more appreciated.

Discriminant analysis

Discriminant analysis was performed to analyze which scales best differentiate between male and female speech, and, separately for men and women, between untrained and trained speech. The statistical procedure was separately performed for men and women, because of the significant interactions between grouping variables Gender and vocal training on several scales (see table 5). Estimated age was not introduced in the analysis, because it is not a bipolar semantic scale.

The discriminant analysis on the scale means per speaker determined the scales "weak-powerful", "monotonous-melodious", "shrill-warm", "creaky-not creaky", "slurring-sprightly", "unpleasant-pleasant", "breathy-not breathy", "dull-clear", "broad-cultured" and "high-low for a (wo)man", in order of importance, to be the set of scales discriminating best between male and female speakers. The canonical correlation of the discriminant function was 0.78, which means that $(0.78)^2 \times 100 = 61\%$ of the variation in the discriminant function is explained by the groups. The percentage correctly classified in their own group on the basis of the classification function coefficients was 90.6%. Sixteen women and five men were incorrectly classified. The centroids (group means) of the two groups on the canonical discriminant function were -1.00 for women and 1.52 for men.

For the female subgroup the discriminant analysis on the scale means per speaker determined the scales "slovenly-polished", "shrill-warm", "creaky-not creaky", "breathy-not breathy", "dull-clear", "monotonous-melodious", "unpleasant-pleasant", "broad-cultured" and "slurring-sprightly", in order of importance, to be the set of scales discriminating best between untrained and trained speakers. The canonical correlation of the discriminant function was 0.63, which means that $(0.63)^2 \times 100 = 40\%$ of the variation in the discriminant function is explained by the groups. The percentage correctly classified in their own group on the basis of the classification function coefficients was 80.7%. Nineteen untrained and seven trained women were incorrectly classified. The centroids (group means) of the two groups on the canonical discriminant function were -0.54 for untrained and 1.20 for trained women.

For the male subgroup the discriminant analysis on the scale means per speaker determined the scales "broad-cultured", "slovenly-polished", "expressionless-expressive", "slurring-sprightly", "high-low", "ugly-beautiful", "unpleasant-pleasant", "dull-clear" and "shrill-warm", in order of importance, to be the set of scales discriminating best between untrained and trained speakers. The canonical correlation of the discriminant function was 0.69, which means that $(0.69)^2 \times 100 = 48\%$ of the variation in the discriminant function is explained by the groups. The percentage correctly classified in their own group on the basis of the classification function coefficients was 80.9%.

Seven untrained and nine trained men were incorrectly classified. The centroids (group means) of the two groups on the canonical discriminant function were -0.91 for untrained and 0.98 for trained men.

DISCUSSION

The results of the conducted experiment show the possibilities of evaluating speech of groups of subjects. Speech of trained subjects was used to give direction to what might be regarded as more "ideal". Untrained subjects without vocal complaints or visually observed abnormalities of the vocal folds produced the speech that was used to establish a frame of reference, consisting of averaged scores on the utilized scales. This frame of reference, which can be acknowledged as representing "normal" speech, is needed to have an image of a "normal" arrangement of perceptual dimensions. Evaluation of speech can be performed by comparing perceptual dimensions with the "normal" arrangement and to specify deviations. To facilitate this evaluation and to offer material for new investigations, speech samples used in this experiment, as well as perceptual specifications and demographic descriptions of each subject were made available on CD-ROMs (SPEX).

Scale properties

The scaling instrument used in this study to perceptually evaluate speech of groups of subjects, showed its practicability and gave proof of its careful construction. Listeners used scale ends with a same representation of a perceptual dimension. Reliably scoring of the scales required a limited number of judges (< 20). Only the scales "creaky-not creaky", "dull-clear" and "slovenly-polished" can be considered as exceptions. The first of these scales "creaky-not creaky" was introduced in this study as a new scale, because of the potential influence of this modality on the perceptive quality of speech. Creaky voice is present in normal speech and regarded as a separate mode of phonation. However, the use of this mode of phonation might differ between groups and, thus, have an influence on --overall-- ratings of speech quality. Although factor analysis of the scores on this scale showed that listeners made judgments as one group, the eigenvalue was lower compared to the other scales, indicating the difficulty that some listeners might have had in giving concise judgments. The same problem might have been present while giving ratings on the other two scales with less effective --however, still high enough-- reliabilities, "dull-clear" and "slovenly-polished". Fagel et al. (8) found changing opinions in judgments on the scale "broad-cultured". It could be that listeners differ in their tolerance regarding an other aspect of articulation as expressed in the scale "slovenly-polished". Compared to previous

investigations (9,12,18) no problems were experienced while dealing with the scale "breathy-not breathy". A high effective reliability of 0.937 was found and factor analysis of the scores on this scale yielded an eigenvalue of 11.5, which explained 42.6% of the variance in the scores. The almost exclusive presence of female listeners might explain the better characteristics of the scale "breathy-not breathy" in this study, as women are known to be more associative raters with higher correlations between scales (18). The higher number of subjects incorporated in this experiment might also have presented a larger variety of breathiness in the stimulus material, thus producing a higher reliability.

Intrajudge reliability was sufficiently high, considering the high values for correlation coefficients (c.c.>0.90) with low probabilities ($p<0.001$). The only exception was presented by the scale "slovenly-polished" (c.c. 0.69, $p=0.028$). It, again, reflects the potential difficulty in rating this articulatory characteristic of speech.

The scale estimated age showed a very high inter- and intrajudge reliability, and there was a high level of agreement among listeners about the use of this scale, considering the eigenvalue of 22.1 of the first factor, explaining 81.9% of the variation in the data. Previous work already established the ability of listeners to adequately estimate age of speakers (9,19-21).

The preparation of the scaling instrument with aligning polarities probably resulted in a more practicable instrument, as listeners can more easily express their qualitative impression of aspects of speech on one side of the rating form, without having to check the correct direction of the polarity. With this aligning an improved version of the original scaling instrument is given for evaluation of laryngeal speech.

Group differences in mean scores

Many differences in scale ratings were found between men and women, as well as between untrained and trained subjects. Speech of women is characterized by more positively judged intonation features, having higher ratings on the scales "expressionless-expressive" and "monotonous-melodious". Intonation is determined for an important part by regulation and variation of the fundamental frequency (F_0) (22). Slow variation of F_0 during an utterance is especially controlled by subglottal pressure, while variation of intralaryngeal muscular activity provide local F_0 movements (22). With these considerations, women should show more variation in activity of intralaryngeal musculature during speech, compared to men. Another positive aspect of female speech is the clearer impression listeners get, which is probably caused by the acoustic characteristics of the smaller dimensions of the female vocal tract (23) and the higher F_0 of women (9).

Although the suffix "for a (wo)man" was especially added to the scale "high-low" to compensate for gender-specific differences in F_0, women were still given a significantly more negative rating on the scale "high-low". It seems that the listeners in this study were not able to compare pitch of the speaker with a gender-neutral image in this specific perceptual dimension and that pitch of men and women was rated systematically to low and high, respectively. The scale "shrill-warm" is closely related to the scale "high-low" and therefore it was no surprise that women were also rated significantly more shrill. Kreiman et al., (5) found a specific perceptual relation between rated degree of pathology and F_0. The observed differences in ratings on the scales "high-low" and "shrill-warm" could thus have an influence on judgments on the scales "ugly-beautiful" and "unpleasant-pleasant" with women having more negatively rated speech.

Compared to speech of men, speech of women was rated softer and weaker, which is in agreement with Awan (24), who measured intensity level of conversational speech of men and women and found significantly softer intensities for female speakers. A second explanation for this difference might be found in the so-called Frequency Code, which suggests that listeners perceive female speech as "small" (25). Louder and more powerful male speech might probably result in a higher intelligibility. However, women might compensate this by a more careful and correct pronunciation (26-28). In the present study only speech of trained women was rated more polished than the male counterpart, which does not give hard evidence for a higher appreciated articulation in females by the group of listeners.

Breathiness is inversely related to glottal closure (29). Speech of untrained females was rated more breathy, which is, therefore, in concordance with previously published results from our research project showing the relatively higher leakage of air (10), as well as the smaller percentage of vocal fold closure in women (3), as compared to men.

Profession and education are known to have an influence on articulation ratings (9). Ratings on the scale "broad-cultured" might, therefore, be influenced by the social background of the subjects. Nearly all untrained male and female subjects were university students or receiving vocational training, respectively, while trained subjects were recruited from choirs with a more diverse social stratification.

In her study, Tielen (9) used an almost identical scaling instrument to compare speech of untrained men and women. Our results are in general in agreement with her study, however a few differences are apparent. In the present study the male speakers are the louder and more powerful ones and women were rated more breathy. Regarding tempo of speech (scales "dragging-brisk" and "slow-quick" in the Tielen study), a same tendency was found with

more positively ratings on the scale "slurring-sprightly" in women. In the Tielen study the effect of an interaction between gender and profession of speaker on the scales "ugly-beautiful" and "unpleasant-pleasant" might have prevented showing difference between men and women in these scales, a difference that can be found in the present study.

Trained speech was rated more expressive and melodious. Phonatory motor control (11) in the trained groups may have provided the subjects in these groups with better intonation abilities. Trained subjects do also have larger intensity ranges (1), which they might employ more fully during speech. The clearer speech of trained subjects might be based on differences in frequency spectrum of the voice source. Compared to untrained subjects typical spectra of singers have a relatively smaller decay in intensity level with higher harmonics and show clustering of formants, producing a so-called singers formant (30). Spectra with more information in the higher frequency region are perceptually characterized as less breathy and more sonorous (31). The changed aspect of trained spectra is related to an increased glottal closure (31). Though not statistically significant, trained subjects in our study had a higher percentage glottal closure than untrained ones (2).

Age was related to a number of scales. Older subjects have speech that is rated warmer on the one hand and more dull on the other. Both characteristics probably depend on ageing of vocal fold structures (2). The more positively rated pleasantness with age in this study is in contradiction with the finding of Tielen (9). A cause might be selection bias, as the average age of trained subjects was older than that of the untrained ones, and speech of trained subjects was rated more positively. Older subjects also received a higher rating on creaky voice, which is often associated with senescence.

Listeners were able to give a good estimation of the age of speakers. Studies showed that estimation might be based on pitch information (32-34) and reading performances (19,34,35).

Factor loadings

Factor analysis resulted in a solution with six factors labelled Intonation, Quality, Physiology, Dynamics, Articulation and Estimated age. Together they explained 93% of the variation in the scores on the scales. The most important factor in evaluating speech was Intonation. Listeners seem to have an attentive ear regarding the perception of variation of speaking fundamental frequency and speaking intensity level (scales "monotonous-melodious" and "expressionless-expressive"), as well as the ability to register the speed of variation of these variables (scales "expressionless-expressive" and "slurring-sprightly"). Positive ratings on these scales are correlated with more pleasant rated speech. Even more important for speech to be rated beautiful and pleasant

are a warm and relatively low voice. Scales representing these perceptual dimensions are clustered in the second factor, Quality. The scale "ugly-beautiful" is also related to three other scales clustered in the third factor Physiology, that is, clear speech without breathiness and creaky voice is rated more beautiful. The fifth factor, Articulation, is the first one with an eigenvalue less than 1. This threshold is normally used to designate the number of factors in a solution. However, because the study of Fagel et al. (8) showed a solution with five factors, each representing a specific perceptual dimension, this threshold was not used in the present study and statistical analysis was forced to produce six factors, one of these separately designated for estimated age. Table 6 shows that cultured and polished articulation is also associated with speech that is rated as being beautiful and pleasant. Estimated age is associated with polished, though broad, speech with a creaky and low voice (see Table 6).

With their data on laryngeal and alaryngeal speech Fagel et al. (8), Nieboer et al. (12) and Tielen (9) followed the same design in evaluating perceptual scores. Typically constructed for the evaluation of alaryngeal speech, Nieboer et al. (12) introduced new scales and left out others, resulting in a factorial solution that is hard to compare with the solution of the present study. In the other two studies comparable factorial solutions emerged. However, labelling of the factors among the studies varied, due to minor differences in scale composition and relative contribution of scales to the specific factors.

Group differences in factor scores

Table 7 presents data on perception of speech in a more comprehensive way by giving means of factor scores for each group. Significance of differences in scores between groups are given in Table 8. The grouping variable gender has a significant effect on Intonation, Quality, Dynamics and Articulation, expressing the large differences that can be obtained while perceptually evaluating speech of men and women. Although differences were found on scale level when comparing speech of trained and untrained subjects, differences were less explicit using factor scores. A significant effect of the grouping variable vocal training was found only on Articulation.
An encouraging aspect of ageing is the higher appreciation listeners have for speech of older persons, regarding the positive relation between age and Quality.

Discriminant analysis

A high percentage of 91% of the subjects were correctly classified for gender. Scales used for this classification come from several factors. The most important scale is "weak-powerful", which refers to the Frequency Code of Ohala (25), suggesting that "size" of male speech is perceived as "large", due to

its lower pitch. The second scale in the discriminant function is "monotonous-melodious", referring to the more emotional impression that listeners have of female speech (9). Scales from the factor Physiology are also important for a correct classification: "creaky-not creaky", "breathy-not breathy" and "dull-clear" have a place in the discriminant function. It points to the difference in vocal function between men and women (10).

In the female subgroup 81% of the subjects were correctly classified for trained or untrained status. Articulation and intonation are processes that can be regulated actively for an important part. These processes, categorized as two separate factors in the present study, are perceptively represented by the scales "slovenly-polished", "monotonous-melodious", "broad-cultured" and "slurring-sprightly" in the discriminant function. The more positive ratings that were given to speech of trained women and the use of these scales for a correct classification stress the presumably more precise articulation, as well as the higher appreciated variation of the variables pitch and intensity in trained women. These qualities of trained women might be based on the experienced level of motor control over the vocal and articulatory organs (1,11). However, a biasing influence of age might be present, considering the positive relation between age and the scale "slovenly-polished". A process merely beyond active control concerns glottal function. A correct classification is also based on specific ratings on the scales clustered in the factor Physiology, trained women showing the more positive ratings. It suggests a difference in vocal fold function between trained and untrained women; however, a related study concentrating on physiology of phonation did not show clear differences between untrained and trained women (10).

A correct classification of 81% was also found for vocal training in men. Important scales for this classification are clustered in the factors Intonation and Articulation, which might both be positively influenced by vocal training, as discussed in the previous paragraph. Physiology seems to be of less importance for a correct classification in the male subgroup.

CONCLUSIONS

With the information given in the previous sections the questions given in the introduction can be answered.

1. Speech of a large number of voice healthy subjects can reliably be evaluated with a scaling instrument such as suggested by Fagel et al. (8).

2. Factor analysis resulted in a solution with six factors, labelled Intonation, Quality, Physiology, Dynamics, Articulation and Estimated age, which represent diverse aspects of speech. With these factors further evaluation of perceptive data on speech can be expedited.

3. Many significant differences in the perception of speech of men and women were found. Speech of women was rated more expressive, melodious, breathy and shrill, higher and clearer on the one hand, and more unpleasant, softer and weaker on the other. Regarding vocal training, speech of trained subjects was rated more expressive, melodious, and pleasant, and also clearer.

4. Significant differences between men and women were also found on factor level. Intonation of women was judged more positively by listeners, whereas more positive ratings were given on Quality and Dynamics of speech of men. On factor level no significant differences were found between speakers with and without vocal training. Social background and education level of subjects demonstrated to have an influence on Articulation.

5. With discriminant analysis scales were selected that can best be used to classify subjects in male and female, and trained and untrained groups. For classification of gender scales were selected from the factors Dynamics ("weak-powerful"), Intonation ("monotonous-melodious" and "slurring-sprightly"), Quality ("shrill-warm", "unpleasant-pleasant" and "high-low for a [wo]man") and Physiology ("creaky-not creaky", "breathy-not breathy" and "dull-clear"). Regarding classification for vocal training, in the female subgroup scales were selected from the factors Articulation ("slovenly-polished" and "broad-cultured"), Quality ("shrill-warm" and "unpleasant-pleasant"), Physiology ("creaky-not creaky", "breathy-not breathy" and "dull-clear") and Intonation ("monotonous-melodious" and "slurring-sprightly"), whereas in the male subgroup scales were selected from the factors Articulation ("broad-cultured" and "slovenly-polished"), Intonation ("expressionless-expressive" and "slurring-sprightly") and Quality ("high-low", "ugly-beautiful", "unpleasant-pleasant" and "shrill-warm").

References

1. Sulter A, Schutte H, Miller D. Differences in phonetogram features between male and female subjects with and without vocal training. J Voice 1995;9:363-77.
2. Sulter A, Schutte H, Miller D. Standardized laryngeal videostroboscopic rating: Differences between untrained and trained male and female subjects, and effects of varying sound intensity, fundamental frequency and age. J Voice 1996;10:175-89.
3. Sulter A, Albers F. (in press). The effects of frequency and intensity level on glottal closure in normal subjects. Clin Otolaryngol 1996:in press.
4. Fex S. Perceptual evaluation. J Voice 1992;6:155-8.
5. Kreiman J, Gerratt B, Precoda K. Listener experience and perception of voice quality. J. Speech Hear Res 1990;33:103-15.
6. Kreiman J, Gerratt B, Precoda K, Berke G. Individual differences in voice quality perception. J. Speech Hear Res 1992;35:512-20.
7. Kuwabara H, Ohgushi K. Experiments on voice qualities of vowels in males and females and in correlation with acoustic features. Lang Speech 1984;27:135-45.
8. Fagel W, Van Herpt L, Boves L. Analysis of the perceptual qualities of Dutch speakers' voice and

pronunciation. Speech Comm 1983;2:315-26.
9. Tielen M. Male and female speech. An experimental study of sex-related voice and pronunciation characteristics. Thesis, Universiteit van Amsterdam, 1992.
10. Sulter A, Wit H. Glottal volume velocity waveform characteristics in subjects with and without vocal training, and relations with gender, sound intensity, fundamental frequency and age. J. Acoust Soc Am 1996:in press.
11. Murray T, Caligiuri M. Phonatory and nonphonatory motor control in singers. J Voice 1989;3:257-63.
12. Nieboer G, De Graaf T, Schutte H. Esophageal voice quality judgements by means of the semantic differential. J Phonetics 1988;16:417-36.
13. Osgood C, Suci G, Tannenbaum P. The measurement of meaning Urbana, IL: University of Illinois Press, 1957.
14. Kim J. Factor analysis. In: Nie H, et al. eds. SPSS: Statistical package for the social sciences. New York: McGraw-Hill Book Company, 1975:468-514.
15. Hatch E, Farhady H. Research design and statistics for applied linguistics. Rowley, London, Tokyo: Newbury House Publishers, 1982.
16. SPSS Inc. SPSSX user's guide (2nd ed.). New York: McGraw-Hill Book Company, 1986.
17. Nunnally J. Psychometric theory (2nd ed.). New York: McGraw-Hill Book Company, 1978.
18. Van Herpt L. Do men and women use a common semantic factor space to describe voice and pronunciation? Proceedings of the Institute of Phonetic Sciences, University of Amsterdam 1987;11:15-25.
19. Ptacek P, Sander E. Age recognition from voice. J Speech Hear Res 1966;9:273-7.
20. Shipp T, Hollien H. Perception of the aging male voice. J Speech Hear Res 1969;12:704-10.
21. Hartman D, Danhauer J. Perceptual features of speech for males in four perceived age decades. J Acoust Soc Am 1976;55:713-5.
22. Strik H, Boves L. A physiological model of intonation. Proceedings Dept. of Language and Speech, University of Nijmegen 1992/1993;16/17:96-105.
23. Fant G. Non-uniform vowel normalization STL-QPSR 1975;2-3:1-19.
24. Awan S. Superimposition of speaking voice characteristics and phonetograms in untrained and trained vocal groups. J Voice 1993;7:30-7.
25. Ohala J. Cross-language use of pitch: an ethological view. Phonetica 1983;40:1-18.
26. Smith P. Sex markers in speech. In: Scherer K, Giles H, eds. Social markers in speech 1979:109-46.
27. Koopmans-van Beinum F. Vowel contrast reduction. An acoustic and perceptual study of Dutch vowels in various speech conditions. Thesis. Universiteit van Amsterdam, 1980.
28. Henton C. A comparative study of phonetic sex-specific differences across languages. Thesis, Oxford, Great Britain, 1985.
29. Södersten M, Lindestad P. Glottal closure and perceived breathiness during phonation in normally speaking subjects. J Speech Hear Res 1990;33:601-11.
30. Sundberg J. The acoustics of the singing voice. Sci Am 1977;236:82-91.
31. Södersten M, Hammarberg B. Effects of voice training in normal-speaking women Videostroboscopic, perceptual, and acoustic characteristics. Scand J Log Phoniatr 1993;18:33-42.
32. Hollien H, Shipp T. Speaking fundamental frequency and chronological age in males. J. Speech Hear Res 1972;15:155-9.
33. Jacques R, Rastatter M. Recognition of speaker age from selected acoustic features as perceived by normal young and older listeners. Folia Phoniatr 1990;42:118-24.
34. Shipp T, Qi Y, Huntley R, Hollien H. Acoustic and temporal correlates of perceived age. J. Voice 1992;6:211-6.
35. Brown Jr. W, Morris R, Michel J. Vocal jitter in young adult and aged female voices. J. Voice 1989;3:113-9.